História da Linguística, Resumida

Para estudantes do processamento da língua natural

por Bruno Loff

Pequena Introdução

O pensamento humano tem um poder gigante. A oportunidade, que surge com a civilização, de não passar o dia a caçar ou a cultivar e de poder exercitar a mente, é a maior das oportunidades. Isto torna-se particularmente claro ao estudar a história da ciência e do conhecimento, e a história da linguística em particular. Parte nenhuma desta viagem do saber seria possível sem a linguagem, e como ser plurifascinado que é o Ser Humano, este debruçou-se também no estudo do fenómeno da linguagem.

Para dar a conhecer um pouco da história deste estudo, faz-se neste texto uma abordagem temporal: Divide-se a história da linguística em quatro fases importantes e temporalmente delimitadas, que correspondem directamente aos capítulos do texto. Conta-se primeiro do nascimento da linguística e do seu desenvolvimento basilar durante a antiguidade. Fala-se de seguida dos desenvolvimentos mais sérios que ocorreram durante o Renascimento e até meados do século XX. O último capítulo relata o avanço fundamental do estudo da linguística computacional que ocorreu neste último século.

Incluído neste texto segue um apêndice que pretende despertar o interesse do leitor para algumas questões fundamentais e filosóficas sobre a linguagem.

A Antiguidade

O mais antigo conhecido curioso sobre os problemas da língua foi um faraó egípcio do século VII A.C. chamado Psamético. Está observado historicamente que os problemas e questões que surgem primeiro àqueles que pioneiramente reflectem sobre a linguagem são sempre semelhantes. As primeiras perguntas que o investigador da linguística virgem fazia eram um resultado do surgimento da consciência da linguagem. Até a um certo momento na história as pessoas falavam e era tudo. A certa altura, pelo menos desde Psamético, o fenómeno torna-se consciente e a primeira pergunta que nasce desta consciência é invariavelmente "De onde veio a linguagem?"

Psamético acreditava que havia uma lingua original, que as pessoas sabiam à nascença. Num espírito muito científico, e sem qualquer entrave moral da sociedade científica da época, Psamético resolveu isolar duas crianças, e ao que parece estes dois seres proferiram algures na sua vida umas poucas palavras de Frígico, língua da Frígia, uma região da Ásia Menor.

Isto satisfez Psamético, que acreditou resolvida a sua questão[1]. No entanto, sendo sempre surpreendente, a história da linguística não foi sempre tão macabra. Foi no séc. V A.C. que um gramático indiano chamado Panini[2] culminou uma série de estudos sobre o Sânscrito. A palavra Sânscrito significa perfeição, ou completude, e o Sânscrito era considerada uma língua sagrada, a língua dos deuses. De facto é uma ideia tentadora. O mais impressionante do trabalho de Panini é a sua gramática. Panini definiu o que se poderia hoje chamar uma álgebra da língua. Escreveu nesta *álgebra* uma gramática com três mil novecentas e cinquenta e nove regras. Esta gramática descreve semântica, sintáctica e morfologicamente o Sânscrito, e é considerada não só uma das mais completas gramáticas escritas para qualquer língua, como uma das mais curtas.

1. É curioso que o rei James V da Escócia fez a mesma experiência vinte e dois séculos depois, chegando a resultados semelhantes, mas as crianças falaram Hebraico.

2. Cujo nome verdadeiramente bem escrito não me está acessível neste pacote tipográfico.

Houvesse uma comunidade cientifica internacional na altura! O trabalho de Panini, a profundidade e o rigor do seu saber, o salto intelectual – quando comparado com qualquer outro estudo da estrutura da linguagem na altura – feito por este senhor só seria vingado e retomado vinte e tal séculos depois, com a invenção das gramáticas generativas.

Mas lá iremos.

Se há homens que dão saltos de gigante, regra geral a história faz-se aos passos. Outra espantosa civilização, os povos Helénicos, iniciaram também eles o estudo da linguagem. Sobre a origem da linguagem há pontos de vista divergentes de diferentes estudiosos da Grécia. Enquanto Platão acreditava, como está explícito no seu diálogo *Crátilo*, que os nomes dados aos objectos do *mundo real* eram-lhes intrínsecos, enquanto que o seu discípulo Aristóteles já pensava que eram apenas uma convenção e que poderiam ser outros quaisquer. É com esta noção que Aristóteles divide as palavras faladas em oito categorias gramaticais. A ideia que deve ser oito o número de categorias gramaticais durará até muito tarde na História. É no século primeiro A.C. que a primeira gramática completa para o Grego é escrita, por Dionísio de Trácia, definindo igualmente oito categorias gramaticais, ligeiramente diferentes das oito definidas por Aristóteles.

Durante os séculos seguintes o povo romano – que herdou, além de muitas outras coisas, a tradição linguística grega – desenvolve gramáticas para o Latim, tendo como base o trabalho de Dionísio de Trácia. Roma teve grandes gramáticos e o Romano grandes gramáticas. Varro, Carísio, Donato, Palaemon, Diomedes, Prisciano... Somando um total de para cima de trinta volumes, tudo sobre o Latim: Uma herança rica e extensíssima nos séculos que se seguiram.

Do Renascimento ao Séc. XX

Em termos de títulos saltou-se a idade média, mas vale talvez a pena dedicar-lhe um parágrafo. Toda a época do século V ao século XV é pobre para o desenvolvimento da linguística. Havia um intenso estudo do conhecimento gramático do Latim, visto ser esta a língua dos estudiosos e da igreja católica, no entanto o aprofundamento de conhecimentos foi ligeiro. É de notar que no principio do séc. XIV Dante escreve *De vulgari eloquentia*, um estudo aprofundado sobre as línguas da época. Naquele tempo a língua erudita era o Latim. Fazia-se ciência em latim, literatura em Latim. No seu livro Dante defende o vernáculo, a língua vulgar, como uma língua igualmente digna.

O Renascimento foi para a linguística, como para a maioria dos ramos do pensamento humano, uma época de crescimento. Foi nos séculos XV a XVII que foram escritas as primeiras gramáticas de diversas línguas – o Inglês, o Francês, e outras – e várias personalidades e instituições começaram um estudo filosófico e analítico da linguagem. Foi nesta época que, na Europa, as noções de "falar bem" começaram a ser disseminadas para outras línguas para além do Latim e do Grego.

Uma grande e franca expansão da linguística nas civilizações ocidentais deu-se nos finais do séc. XVIII. Munidos do extenso conhecimento gramatical de diversas línguas, alguns intelectuais começaram a notar muitas semelhanças entre o Sânscrito, as línguas latinas, o Grego e as línguas germânicas. Foi Sir William Jones que propôs a ideia de todas estas línguas terem um antepassado comum, uma língua que é desde então denominada Indo-europeu, que teria sido falada algures entre a Europa Oriental e o Oriente Médio cerca de 8 mil e 4 mil anos A.C., muito antes da invenção da escrita. Este trabalho desencadeou o denominado movimento histórico-comparativo, ou linguística histórica. O método do histórico-comparativismo apela bastante à razão. Sob a suspeita de que duas linguagens partilham um antepassado comum, o histórico-comparativista procurará nas duas linguagens palavras com significados muito próximos e sons bastante semelhantes e conjecturará depois como seria a língua original. O leigo que escolha perseguir este método nos seus tempos livres chegará, quem sabe?, ao Frígico ou ao Hebraico.

Se o estudo gramatical das línguas nos explica a estrutura da língua e o movimento histórico-comparativista nos ensina as origens das línguas, falta no entanto um ramo do conhecimento linguístico que estude o fenómeno do significado e da comunicação da linguagem – O que é, para que serve e como funciona a linguagem? Como e o que é que se transmite informação com a linguagem? No final do séc. XIX, ao tentar responder a estas questões, Ferdinand de Saussure aplica uma metodologia por ele inventada para abordar o fenómeno da linguagem. Esta metodologia, o estruturalismo, procura no objecto de estudo os elementos fundamentais que o compõem. Foi esta dissecação que Saussure fez estudando a linguagem, iniciando um ramo do saber que chamou Semiologia, hoje denominado de Semiótica. É este senhor que distingue as noções de língua e fala – a língua é um sistema de regras e convenções sociais e a fala é a concretização da língua por um individuo – e chama a atenção para a existência do significado e do significante – em que o significado é o conceito que se quer transmitir e o significante é o meio de transmissão do significado (a escrita, a fala). Estas distinções são úteis e relevantes quando se pensa na linguagem, e em sistemas que lidam com a linguagem.

O próximo grande passo na disciplina da linguística foi o dado por Noam Chomsky. Em 1957, Noam Chomsky escreve o livro *Syntactic Structures*, onde descreve as gramáticas generativas, uma sua invenção. O que Chomsky inventou foi uma ferramenta que permite um estudo formal da linguagem. É este estudo do ponto de vista formal que domina os grandes avanços da linguística durante o resto do século XX.

O aluno de processamento da língua natural está, talvez sem o saber, extremamente familiarizado com as gramáticas generativas de Chomsky. Uma gramática generativa é um instrumento formal para definir completamente um conjunto de linguagens – as linguagens generativas. Uma gramática generativa é um conjunto de regras de transformação de símbolos, que a partir de um símbolo inicial geram todas as sequências de símbolos possíveis de uma dada linguagem. Qualquer linguagem generativa pode, portanto, ser descrita com um numero finito de regras, mesmo na possibilidade da linguagem ser infinita.

Há uma necessidade histórica de explicar a qualquer aluno de processamento de língua natural outro conceito inventado por Chomsky, e que hoje se denomina apropriadamente a hierarquia de Chomsky. Dependendo da forma das regras de uma dada gramática generativa, esta gramática descreve uma linguagem que pode ser classificada numa hierarquia de linguagens, de acordo com a potência computacional de uma máquina que consiga, dada uma sequência de símbolos arbitrária, saber em tempo finito se essa sequência pertence à linguagem, isto é, de acordo com a potência computacional necessária para *reconhecer* a linguagem.

As linguagens podem desta forma ser do tipo 0, 1, 2 ou 3, de acordo com o seu poder expressivo. Todas as linguagens generativas são do tipo 0, e o conjunto das linguagens generativas é exactamente o conjunto de linguagens reconhecíveis por um computador com memória infinita. As linguagens tipo 1, também conhecidas por linguagens com contexto, são aquelas que um computador com uma memória limitada[3] consegue reconhecer. As linguagens do tipo 2 são porventura bastante familiares ao estudante de processamento de língua natural: são as linguagens livres de contexto – reconhecíveis por um autómato de pilha. Por fim, as linguagens tipo 3 são aquelas que podem ser reconhecidas por um autómato finito.

Para compreender melhor os anteriores elogios ao sistema gramatical de Panini, servirá talvez dizer que as linguagens generativas são exactamente as linguagens reconhecíveis por gramáticas que usem o sistema de regras Paniniano. Há inclusive abordagens Paninianas ao estudo do processamento da língua natural[4].

3. De facto a memória é "limitada" por uma função linear do tamanho do input, e não uma memória de tamanho fixo. Isto corresponde à noção intuitiva de que para utilizar um computador para reconhecer sequências de símbolos de tamanho n, então a memória do computador deve poder conter alguns $O(n)$ símbolos para processamento intermédio. É de notar que se for absolutamente necessário que caibam $O(n^2)$ na dita memória, a função já não é linear e portanto a linguagem não é do tipo 1.

4. Ver "Natural Language Processing: A Paninian Perspective" por Bharati et al.

No séc. XX surgiram também diversos ramos da linguística que abordam as questões da linguagem de outras perspectivas: a bio-linguística, a psico-linguística, a socio-linguística e a linguística cognitiva, para nomear alguns. Além destes ramos da linguística, surgiu também no séc. XX a linguística computacional, da qual trataremos de seguida.

História da Linguística Computacional

Com o advento da segunda guerra mundial e da consequente necessidade de desencriptar as comunicações alemãs surgiram diversas frentes de investigação criptográfica. Shannon, Good e Turing inventaram poderosas ferramentas estatísticas para processar informação e linguagem e havia, no fim dos anos quarenta, a ideia que uma língua podia ser vista como uma forma encriptada de outra. Assim sendo, tentou-se nos anos a seguir à guerra uma abordagem estatística à tradução automática, analisando milhares de frases com as ditas ferramentas e tentando extrair dessa forma métodos para tradução. Citando Bar-Hillel: "Um linguista munido de alguns assistentes deverá ser capaz de providenciar um sistema para qualquer linguagem razoavelmente bem descrita – como o inglês, o alemão ou o russo – dentro de um ou dois anos".

Esta promessa é bem mais difícil de cumprir do que Bar-Hillel acreditava, e portanto, depois de muitos milhões investidos, houve no fim dos anos cinquenta um corte no financiamento, e uma desilusão com as promessas dos investigadores de processamento da língua natural.

Foi apenas com o advento da obra de Chomsky que a linguística passaria a ter um papel activo na investigação em processamento da língua natural, não por falta de participação dos linguistas, mas por falta de ferramentas linguísticas que lhes permitissem dar uma contribuição útil. Com o advento dessas ferramentas, os métodos estatísticos passam para segundo plano, e só serão retomados mais tarde.

Dos anos sessenta aos oitenta inventaram-se muitas coisas interessantes. Começou o trabalho que utiliza as gramáticas generativas no processamento da língua e surgem os primeiros sistemas de síntese de fala. O conceito de função temática, as gramáticas semânticas, as redes semânticas, a teoria de Schank de dependência conceptual, a teoria dos enquadramentos, foi tudo inventado nestes anos. Surgem assim diversos sistemas que funcionam excepcionalmente bem em domínios limitados[5], tendo taxas de sucesso acima dos 80%.

Durante os anos oitenta houve uma série de avanços em reconhecimento de fala; os laboratórios da IBM voltaram a pegar nos métodos estatísticos e inventaram a base para a maioria dos métodos de reconhecimento de fala utilizados actualmente. Há um retorno à investigação na tradução automática e inicia-se a construção de diversos recursos linguísticos.

Foi desde os anos noventa até à actualidade que a linguística computacional e o processamento de língua natural fizeram as suas maiores conquistas. O reconhecimento de fala tornou-se uma actualidade prática, a tradução automática existe e está ao alcance de todos, há uma ampla quantidade dos mais variados recursos linguísticos. Todos os campos da linguística computacional e do processamento de língua natural estão em franca expansão.

É uma altura oportuna para trabalhar em processamento de língua natural.

Apêndice: Algumas Questões

A linguagem é utilizada em tudo o que fazemos – pensamos e comunicamos com linguagem. Surge então um truismo: A linguagem é importante. Espera-se que nenhum leitor para quem este ponto não esteja bem assente chegará a ler estas linhas.

5. Por exemplo: SHRDLU, LUNAR, LIFER/LADDER.

Sugere-se, tendo em conta este ponto de vista, que os leitores façam as seguintes experiências conceptuais[6]:

1. Imagine-se um aluno extremamente inteligente, que domina todas as matérias de uma dada disciplina com um grau extremo de profundidade. Infelizmente, o aluno tem graves dificuldades de expressão. Num evento avaliativo da dita disciplina, o aluno faz um projecto e escreve um relatório sobre o projecto. Suponhamos que este evento é tal que a autenticidade da autoria do projecto é garantida: Seguramente que não houve cópia.

   O projecto supera as expectativas do professor: está integralmente correcto e não possui nenhuma falha conceptual. O relatório, no entanto, é um emaranhado tamanho de frases mal construídas e de ideias mal explicadas, todas juntas num documento de estrutura duvidável, cuja leitura e compreensão são difíceis, se não impossíveis.

   O professore deve atribuir uma boa nota ou uma nota medíocre?

   Suponha primeiro que o professor atribuiu boa nota, e depois suponha que atribuiu uma nota medíocre. Que justificação tem o professor em cada um dos casos?

2. Imagine-se um professor genial num certo ramo do saber. Este senhor é tão conceituado entre os seus pares, que estes o consideram como o Grande Homem daquela disciplina, a suma autoridade sobre o assunto – o seu currículo e produção de resultados conferem-lhe uma inegável prova de entendimento profundo das matérias. No entanto, este ilustre personagem não o é para os seus alunos. Possuindo todo aquele saber indisputável, a sua capacidade de o transmitir ao não-técnico é nula. Balbucia em termos demasiado técnicos, ou pretende transmitir noções gerais sobre assuntos dos quais os alunos não dominam a base. É trapalhão: engana-se sistematicamente na sua exposição, e quando corrigido por algum aluno admite a falta, mas dá uma explicação tão complexa para a contornar que o aluno, de si já perdido, exaspera.

   Admira o intelecto deste homem? Porquê e em quê? Como justificaria a opinião contrária à sua?

3. Imagine-se um mundo hipotético em tudo semelhante ao nosso, mas onde as instituições educativas tinham a tradição secular de torturar os seus novos alunos. Neste mundo hipotético tal como no nosso a tortura tornou-se proibida.

   Imagine-se neste mundo uma livre associação de alunos num dado *Louvado e Emeretíssimo Instituto C*, L.E.I.C., cujo objectivo professado é receber os novos alunos, inventando actividades de convívio saudável para mostrar o L.E.I.C. aos alunos recém chegados, explicar-lhes onde são as salas e quem são os professores, e iniciar com esses alunos uma série de relações agradáveis e fraternais.

   Os alunos desta associação decidiram chamá-la de "Comissão de Tortura do L.E.I.C.", ou abreviando, de CTLEIC.

   Inquiridos sobre o estranho nome, que parecia tão dissociado dos objectivos do grupo, os membros defenderam-se dizendo que é apenas um nome, e que a ideia do grupo é fazer actividades de tortura na brincadeira, nada como nos velhos tempos, e que pretendem com tais actividades re-inventar o conceito de tortura na universidade, para que a palavra "tortura" possa ser utilizada sem receio, tornando-se reminiscente de uma actividade agradável e prazenteira.

   Que críticas e que louvores merece a CTLEIC?

   Que atitude previsível tem o aluno recém chegado ao L.E.I.C. quando se depara com este nome? Porque é que ele tem esta atitude? O dito nome funciona a favor da CTLEIC?

6. Como experiências conceptuais, parece-nos terem uma relevância independente do contexto de qualquer leitor; no entanto, todas as ditas experiências são baseadas em acontecimentos reais e bastante próximos do autor.

> Concorda que CTLEIC é “apenas” um nome, ou acha que o nome escolhido é uma coisa importante? Porquê? Que justificações há para a opinião contrária à sua?

Espera-se que estas situações façam o leitor reflectir sobre a importância da linguagem, chegue às conclusões que chegar.